Tumor Ecology and Complementary Information

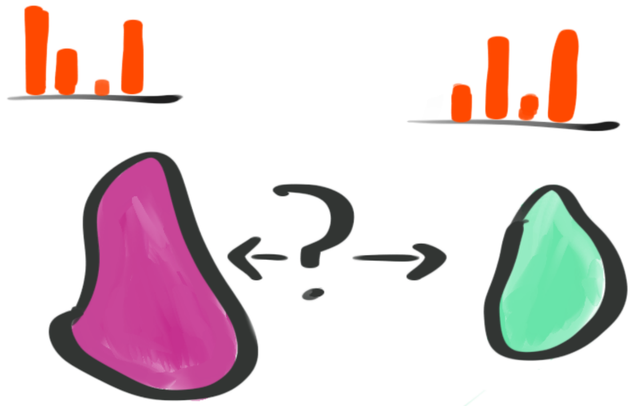
BIRS Integrative Analysis of Emerging Biological Data Types
June 16, 2020

Kris Sankaran
UW Madison Statistics Dept.

Code: github.com/krisrs1128/birs_mini
Vis: https://observablehq.com/@krisrs1128/spatial-vs-expression-map
Binder: https://mybinder.org/v2/gh/krisrs1128/birs_mini/master?urlpath=rstudio
Slides: https://tinyurl.com/yanphfmg

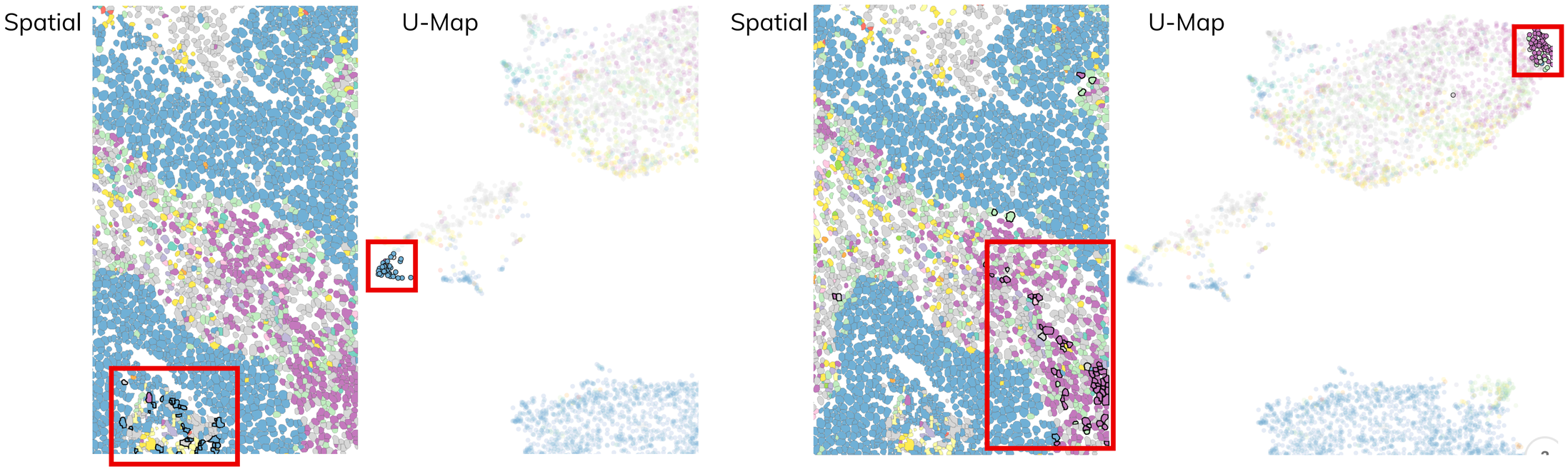
Problem Setting

- Tumor ecosystems
  - We can now study cancers as ecosystems of interacting cells
  - Interactions have consequences for disease progression
- Data sources
  - Mass Spectrometry: Composition of the cells in the ecosystem
  - MIBI-TOF: Interactions between cells



Interactive Visualization

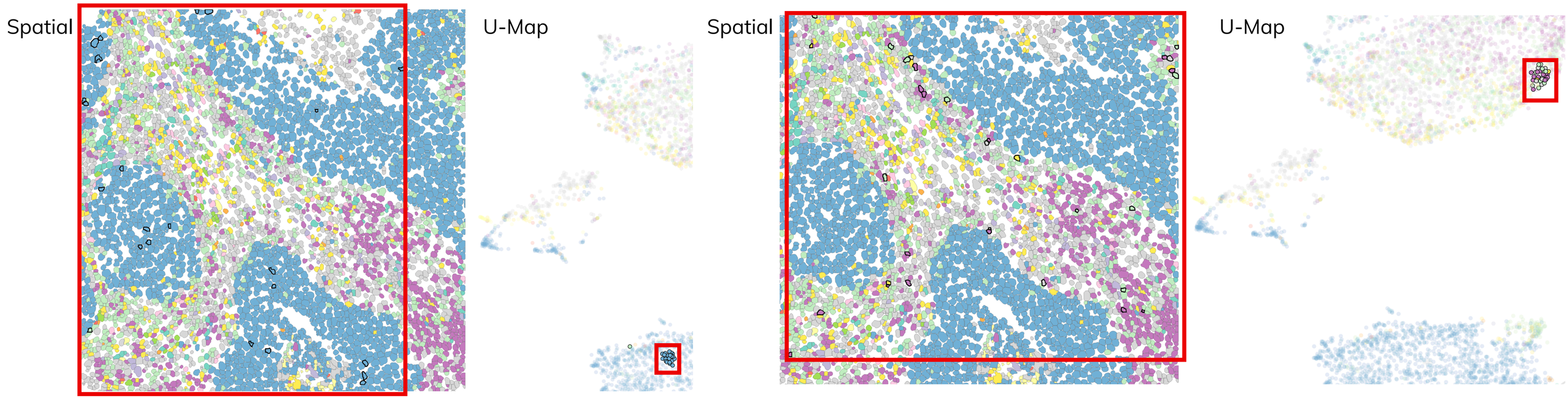
- Linked Brushing: Combine (literal) spatial map with abstract (U-)map
- Within cell-types, some U-Map clusters are spatially co-located, but far from universal



Examples where U-Map clusters correspond to spatially nearby cells. Immune highlighted in left pair, tumor on right.

Interactive Visualization

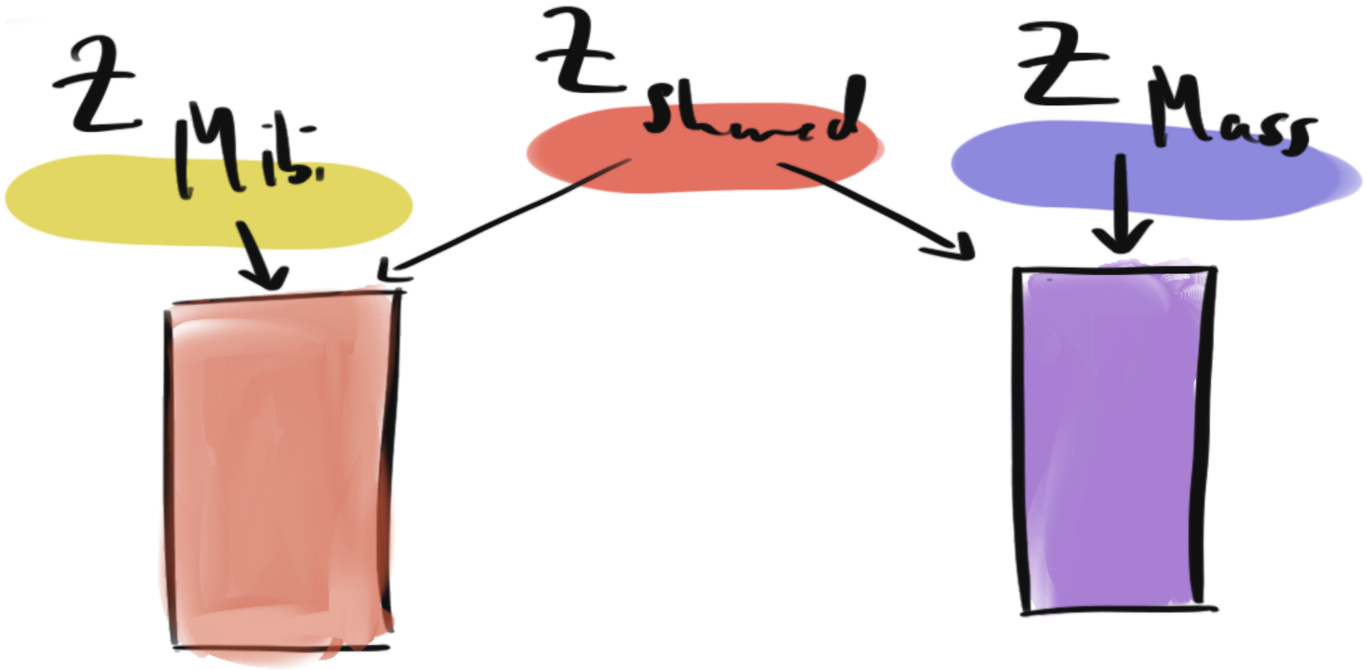
- Linked Brushing: Combine (literal) spatial map with abstract (U-)map
- Within cell-types, some U-Map clusters are spatially co-located, but far from universal



Examples where U-Map clusters are spatially diffuse. Immune cells highlighted in left pair, tumor on right.

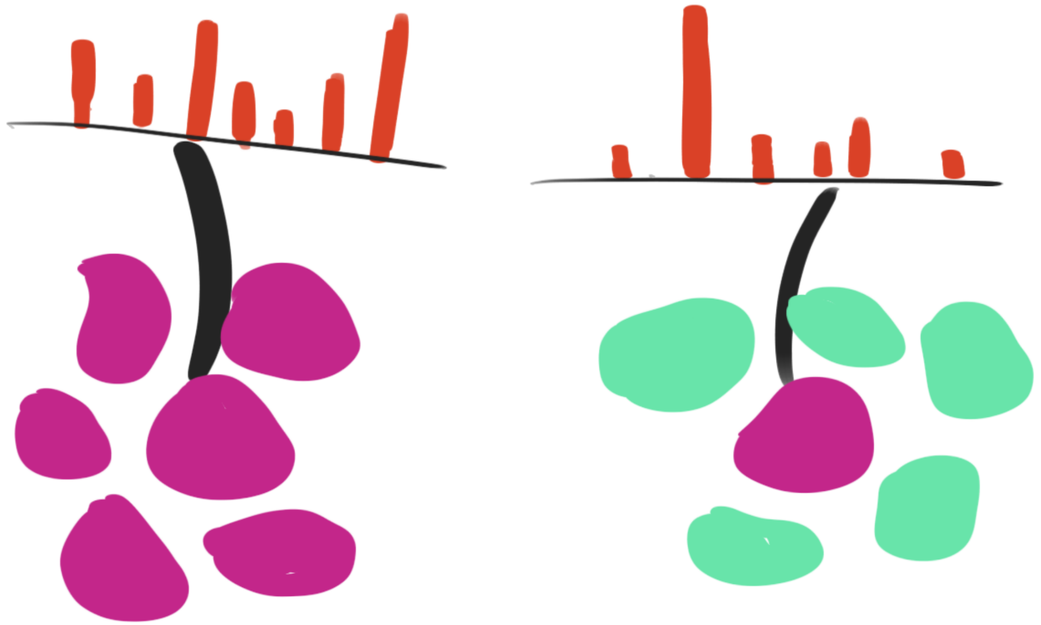
Cell-Level Analysis

- Can we recover shared latent phenomena?
- To what extent can a simple assay be a proxy for a powerful one?
  - Can the trade-offs guide experimental design?



Proposal: Direct inversion

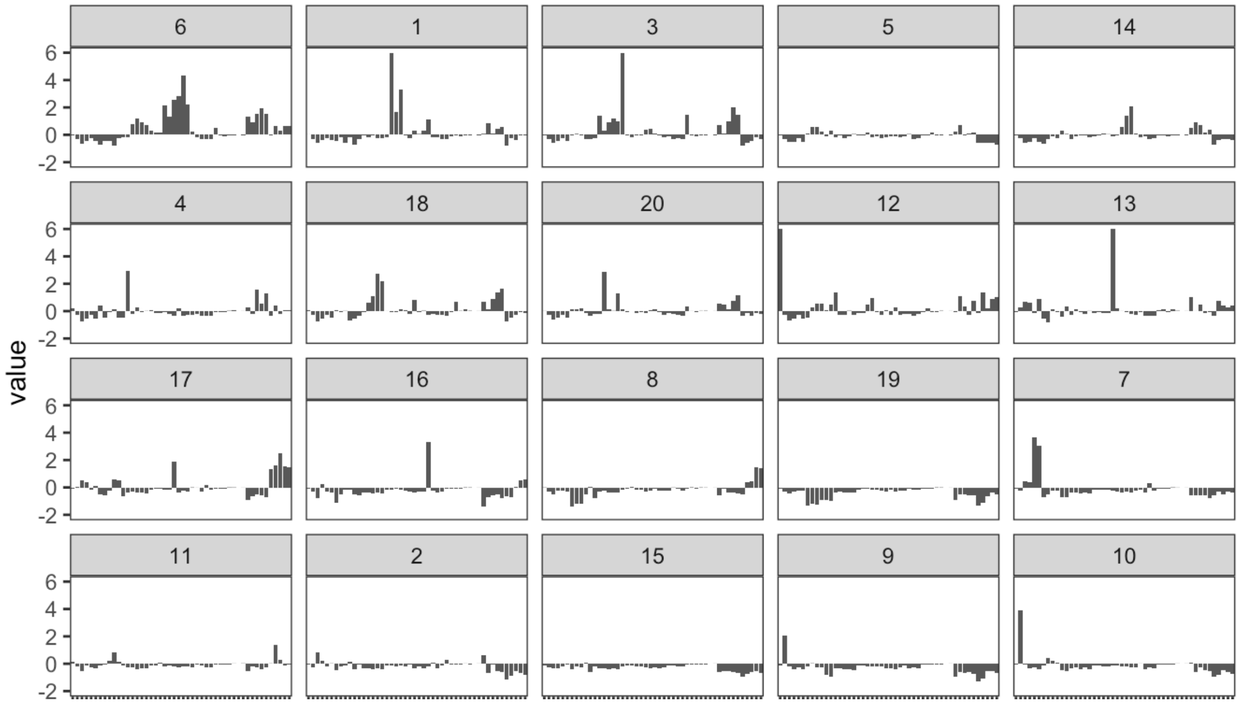
- Rigorous: latent variable analysis, specifying full generative mechanism
- Hack (but simple!): *Train* a protein-to-spatial expression model using MIBI-TOF, and then *test* that on Mass Spec
- Find whether given configurations of neighboring cells force specific expression patterns (especially if configuration is unrelated to simply composition)... by trying to learn the inverse



Proposal: Direct inversion

Recipe,

1. ***Cluster***: Make clusters, from expression data
2. ***Featurize***: Define spatial features
3. ***Embed***: Reduce dimensionally of spatial features
4. ***Predict***: Using expression alone, predict spatial embeddings
    A. Only use proteins available in Mass Spec

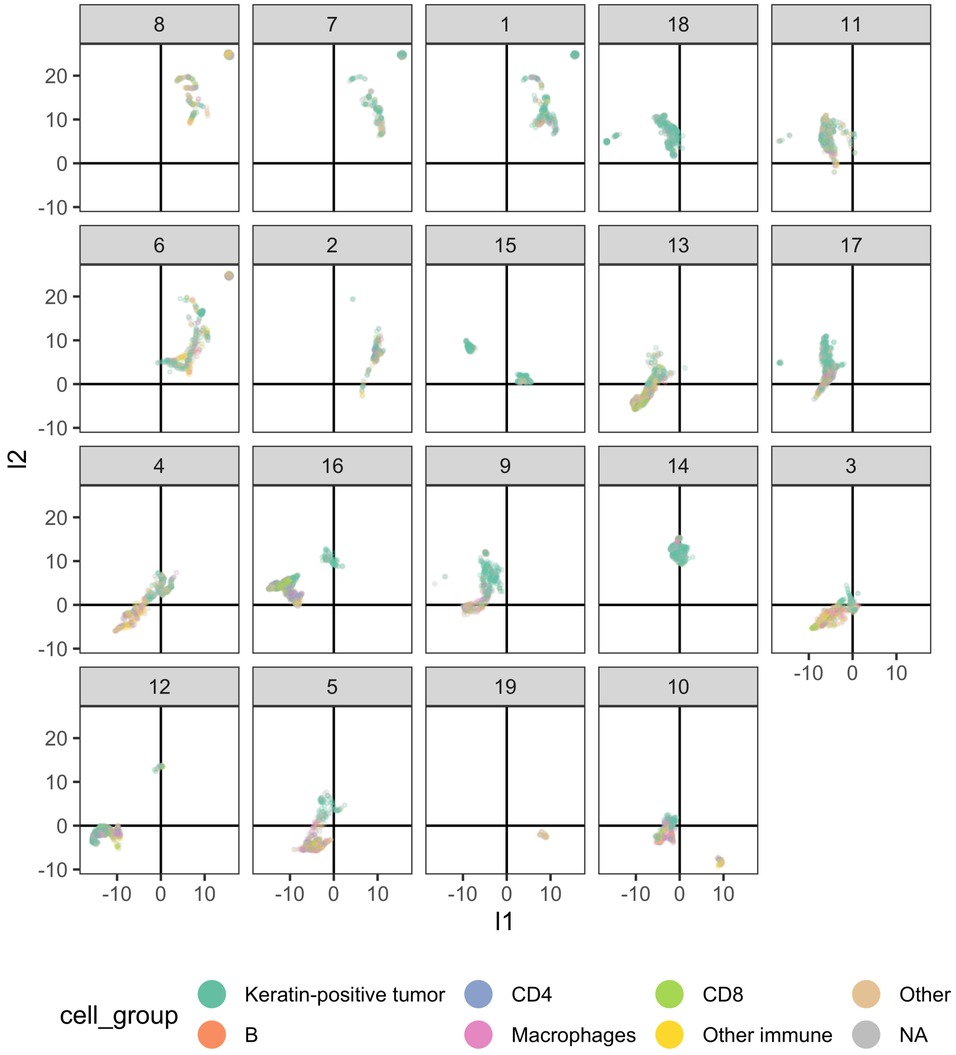


20 centroids from the clustering. Each column is a protein.

Proposal: Direct inversion

Recipe,

1. ***Cluster***: Make clusters, from expression data
2. ***Featurize***: Define spatial features
3. ***Embed***: Reduce dimensionally of spatial features
4. ***Predict***: Using expression alone, predict spatial embeddings
    A. Only use proteins available in Mass Spec



Embeddings of the neighborhood proportion vectors.

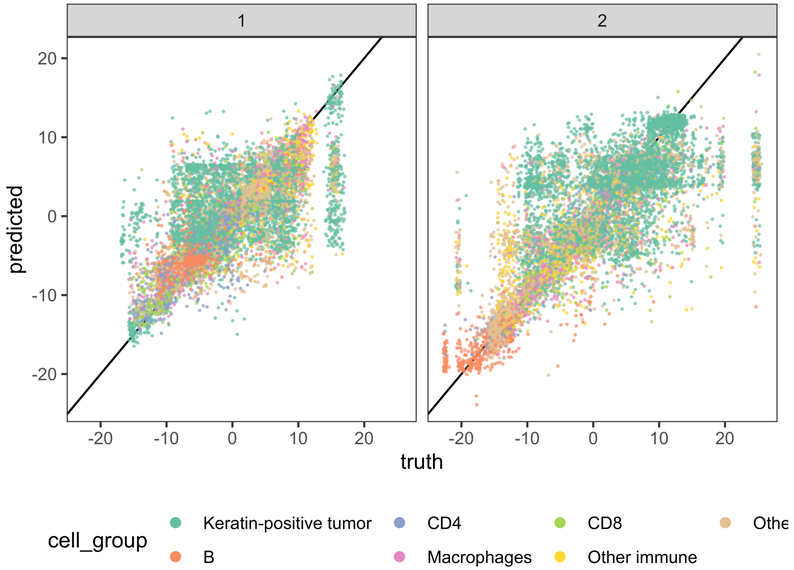
Proposal: Direct inversion

Recipe,

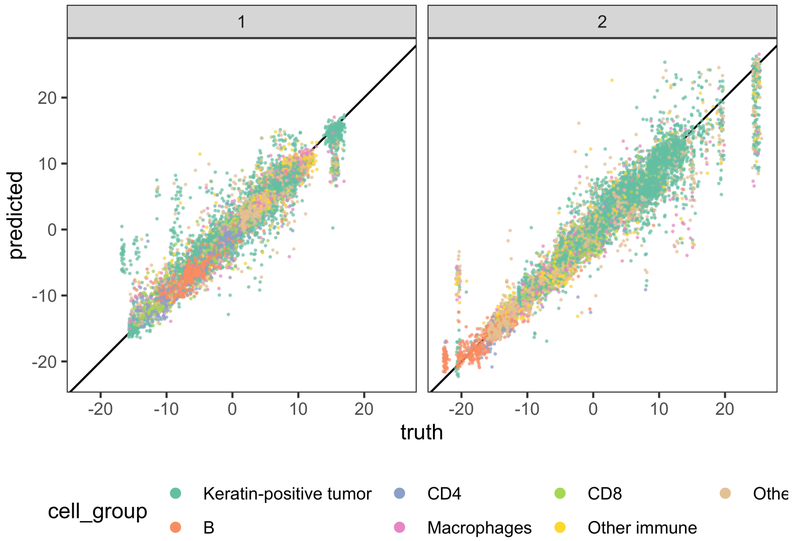
1. ***Cluster***: Make clusters, from expression data
2. ***Featurize***: Define spatial features
3. ***Embed***: Reduce dimensionally of spatial features
4. ***Predict***: Using expression alone, predict spatial embeddings

    A. Only use proteins available in Mass Spec

(a)



(b)



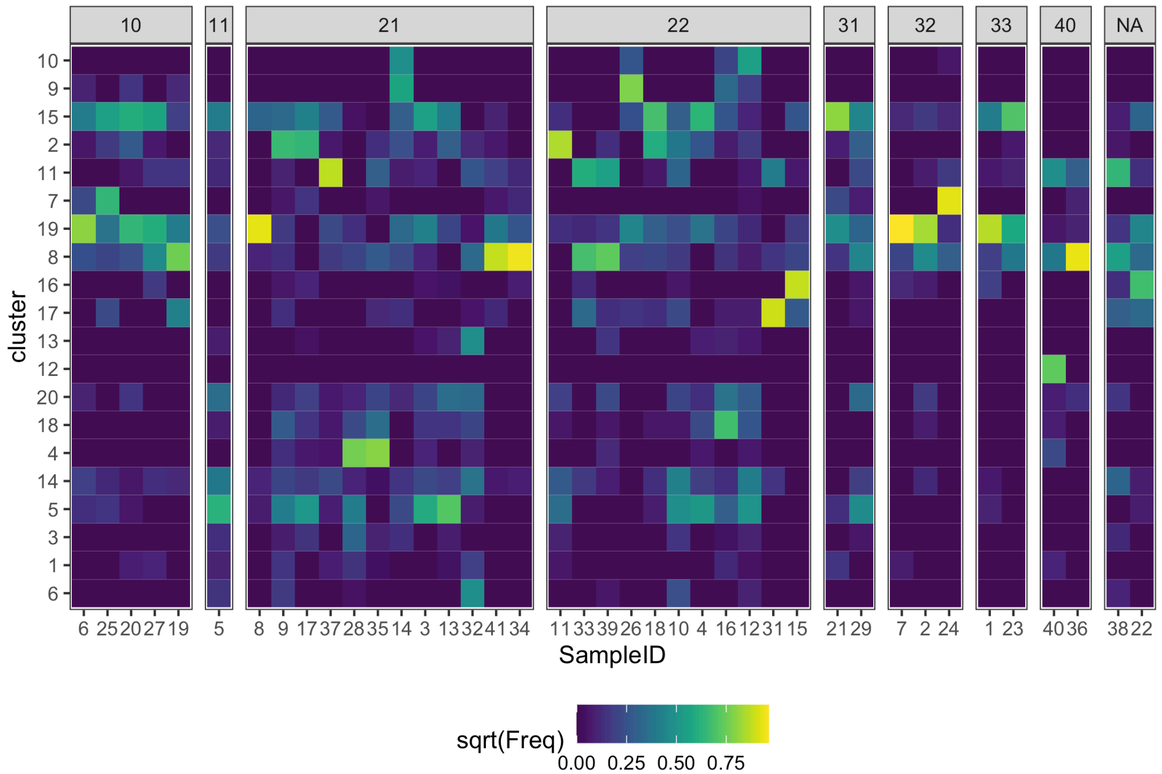
Prediction performance, when using expression data from (a) just mass spec and (b) using all proteins. Two columns are two dimensions of the embedding.

Sample-Level Analysis

- Many scientific claims are about the entire ecosystem, not individual cells
  - E.g., Tumor heterogeneity
- Interaction vs. Composition
  - Interactions between cells might be hard to find
  - Ecosystem properties may be visible from composition alone

Expression -> Spatial (Sample Level)

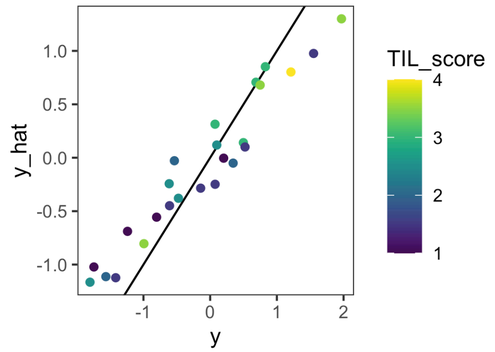
- Recipe,
    1. ***Cluster***: Make clusters, from expression data
    2. ***Featurize***: Define spatial features
    3. ***Aggregate***: Data are at cell level, but we need summaries at sample level. So compute functions of spatial features / find cluster mixing %s.
    4. ***Predict***: Predict spatial features from cluster counts in (2)
- Intuition: I(X, Y) is large if communication channel has low noise

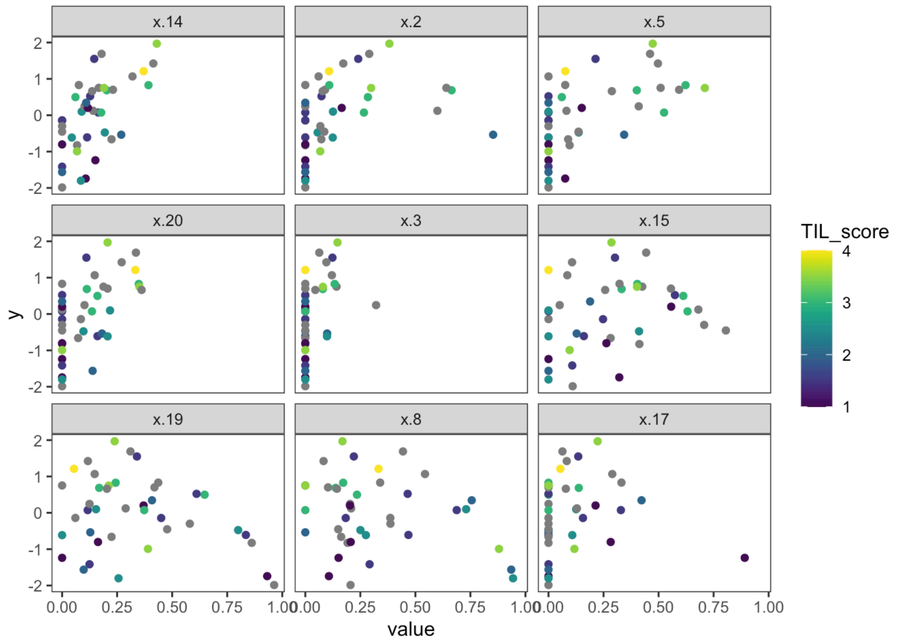


The representation of each sample (column), based on the %s of cells it has from different clusters.

Expression -> Spatial (Sample Level)

- Recipe,
    1. ***Cluster***: Make clusters, from expression data
    2. ***Featurize***: Define spatial features
    3. ***Aggregate***: Data are at cell level, but we need summaries at sample level. So compute functions of spatial features / find cluster mixing %s.
    4. ***Predict***: Predict spatial features from cluster counts in (2)
- Intuition: I(X, Y) is large if communication channel has low noise





Predicting the average cluster entropy of all the 5-nearest neighbor balls within a person, based only on expression data.

Phenotype = Spatial + Composition

- As an alternative measure of redundancy, see how much performance improves when combining two tables
  - In linear regression, adding redundant variable decreases performance
- Approach only works if we have easily predictable phenotypic characteristics

Spatial

mtry	RMSE	Rsquared	MAE
2	0.9747601	0.1128382	0.8176102
3	0.9978172	0.1054846	0.8295354
4	1.0084429	0.1107027	0.8332106

Expression

mtry	RMSE	Rsquared	MAE
2	0.8173901	0.3116003	0.7030003
11	0.8713073	0.1950925	0.7456817
20	0.9029941	0.1718520	0.7682995

Combined

mtry	RMSE	Rsquared	MAE
2	0.7697287	0.3943017	0.6667050
13	0.8187737	0.2448980	0.7056521
24	0.8478903	0.2337036	0.7246190

Ability to predict TIL increases when we include both sets of features, but there is overlap. Caveat: there are only 25 samples with TIL score available.

Takeaways + Next Steps

- The rows and columns of X should not be taken for granted
  - Several definitions of sampling units work (i.i.d. is a construct)
  - The features must be defined (really, should be learned)
- Degree of redundancy, and source-specific signal, are important
  - It would have been if weird spatial patterns were exactly recoverable from Mass Spec
  - Potentially useful meta-tool (wrapping integrative ‘omic algorithms)